

Variable-rate Retransmissions for Incremental Redundancy Hybrid ARQ

Leszek Szczecinski, Ciro Correa[‡], and Luciano Ahumada[‡]

INRS-EMT, Montreal, Canada

[‡]Escuela de Ingeniería Informática, Universidad Diego Portales, Santiago, Chile

leszek@emt.inrs.ca, ciro.correa@mail.udp.cl, luciano.ahumada@mail.udp.cl

Abstract

The throughput achievable in truncated Hybrid ARQ protocol (HARQ) using incremental redundancy (IR) is analyzed when transmitting over a block-fading channel whose state is unknown at the transmitter. We allow the transmission lengths to vary, optimize them efficiently via dynamic programming, and show that such a variable-rate HARQ-IR provides gains with respect to a fixed-rate transmission in terms of increased throughput and decreased average number of transmissions, reducing at the same time the outage probability.

Index Terms

Automatic Repeat Request, ARQ, Hybrid ARQ, HARQ, Incremental Redundancy, IR, Block-fading channel, Throughput

I. INTRODUCTION

Automatic repeat request (ARQ) uses retransmissions to recover data lost due to errors inevitable when transmitting over variable and unreliable channels. ARQ is based on the principle that the receiver can inform the transmitter about the transmission failure, to which the transmitter

This work was supported by the 7th framework program of European Community FP7/2007-2013 under the grant #236068, and by Fondecyt under grant #1095139, and Anillo ACT-53/2010. When this work was submitted for publication, L. Szczecinski was on sabbatical leave with CNRS, Laboratory of Signals and Systems, Gif-sur-Yvette, France. The results were presented in part at IEEE Global Communication Conference, 6-10 Dec. 2010, Miami, USA.

responds retransmitting the lost data; ARQ used together with channel coding is known as hybrid ARQ (HARQ) [1]. HARQ where we limit the number of allowed transmission attempts is known as truncated HARQ.

In this work, we evaluate the throughput achievable in wireless links when using a truncated HARQ that conveys incremental, redundancy (IR) in subsequent transmission attempts. For such a HARQ-IR system, we use random coding and maximum likelihood decoding assumptions of [2] [3] [4]. We adopt the same simple scenario where each transmission attempt is carried out over independently fading channel and we generalize the assumptions of [2] allowing the transmission lengths (or – rates) to vary throughout the transmissions attempts. We show how to efficiently find the throughput-maximizing rates and we show gains obtained for a finite number of transmissions (truncated HARQ).

The idea of using variable-rate transmissions was already proposed and/or discussed in the literature but was not analyzed in the information-theoretic framework of [2], which sets the upper bounds on the performance of any practical scheme. For example, a general formulation of the problem was provided in [5] which analyzed the infinite number of transmission attempts in abstraction of the channel model. The gains of variable-rate transmission over its fixed-rate counterpart for the predefined families of code were shown in [6] [7] [8] [9]. In [10] [11] the correlated fading was considered, while [12] assumed that the channel stays constant for all transmission attempts. The idea of varying the transmission parameters appeared also in [13] [14] [15] [4], where power was varied on a per transmission-attempt basis.

We are interested here in the practical case of truncated HARQ when the packet loss (outage) cannot be avoided. In such a case the throughput of HARQ may be optimized under constraints imposed on the outage probability [10] [15] or without such constraints [13] [6]; the latter approach is also adopted in this paper.

In this work we analyze the “conventional” HARQ, i.e., when the return channel can carry only one-bit ACK/NACK messages [6] [7] [10]. If, on the other hand, we allow the return channel to carry more bits, then, the parameters (rate or power) can be *adapted* using such a “rich” or “multi-level” feedback, e.g., [9] [16] [4]. In the conventional’ case, the *adaptation* is not possible but the transmission parameters (rate or power) can be *allocated*, that is, defined a priori for given channel conditions (e.g., the average SNR); this is focus of this work.

While power *adaptation* improves the throughput [4], the power *allocation* improves the

diversity (asymptotic value of the outage for high SNR) but yields significant gains in terms of throughput only in the low-SNR range [13]. Such conclusions resemble those drawn in the context of adaptive modulation and coding [17] or in information-theoretic analysis of water-filling [18]. In this work, interesting in medium-high SNR region, we assume a constant-power transmission as the gains obtained when allocating the power are often small [13].

The objective of this work is thus to evaluate the benefits of constant-power, variable-rate transmission for truncated HARQ when compared to the fixed-rate case analyzed in [2] and the main contributions are the following: a) we show how to efficiently optimize the rates allocation for truncated HARQ with incremental redundancy, and b) we asses the gains of variable-rate HARQ over its fixed-rate counterpart, showing that larger throughput, lower outage, and smaller average number of transmissions are yield.

II. SYSTEM MODEL

In the transmission system under study, information bits are separated into packets of equal length of N_b bits, which are then encoded into codeword of N_s complex symbols x_1, x_2, \dots, x_{N_s} that are drawn randomly from the zero-mean complex Gaussian distribution with unitary variance. The symbols are gathered into K sub-codewords $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K$ whose respective lengths are $N_{s,1}, N_{s,2}, \dots, N_{s,K}$. We consider two ways of obtaining the sub-codewords:

- 1) A repetition coding (RTC), where the symbols are picked consecutively starting always with x_1

$$\mathbf{x}_k = [x_1, \dots, x_{N_{s,k}}], \quad N_{s,k} \leq N_s. \quad (1)$$

In this way, $\min_k \{N_{s,k}\}$ symbols are the same in the transmission attempts $1, \dots, k$.

- 2) An incremental redundancy (IR) transmission, where each sub-codewords is composed of different symbols

$$\mathbf{x}_k = [x_{t'_k+1}, \dots, x_{t'_k+N_{s,k}}] \quad \text{with} \quad t'_k = \sum_{l=1}^{k-1} N_{s,l} \quad (2)$$

This corresponds to puncturing of the codewords $\mathbf{x} = [x_1, \dots, x_{N_s}]$ into K distinct sub-codewords \mathbf{x}_k each of length $N_{s,k}$, $k = 1, \dots, K$, where $\sum_{k=1}^K N_{s,k} = N_s$ and $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_K]$.

For convenience, we normalize the values of $N_{s,k}$ using $\rho_k = N_{s,k}/N_b$, which has the meaning of the redundancy (measured by the number of channel uses per transmitted bit) and satisfy the

relationship $\rho = N_s/N_b = \sum_{k=1}^K \rho_k$. We define also the rate of each transmission attempt $R_k = 1/\rho_k$ and since the rate of the transmission attempts are not the same, we talk about variable-rate (VR) transmission, while if $\rho_k \equiv \rho_1, \forall k$ (or $R_k \equiv R_1$) we obtain the fixed-rate (FR) transmission considered before in [2] or [3].

The ARQ process for each packet starts sending the sub-codeword \mathbf{x}_1 . We assume that the feedback (or, *return*) error-free channel exists, which allows the receiver to send (to the transmitter) a one-bit message required by the ARQ process (ACK or NACK). If the packet is not decoded correctly¹, the NACK message is communicated by the receiver to the transmitter. Upon reception of a NACK message, knowing that the first sub-codeword was not decoded correctly, the transmitter sends a sub-codeword \mathbf{x}_2 composed of $N_{s,2}$ symbols. After unsuccessful decoding, another NACK message is generated to which the transmitter responds sending the codeword \mathbf{x}_3 . This continues till the maximum allowed number of transmission attempts K is reached (truncated HARQ) or until an ACK message, denoting a successful decoding, is received.

In a particular case of $\rho_k \equiv \rho_1$, the sub-codewords have the same length/rate and we recover the retransmission schemes analyzed in [2].

The channel remains constant during transmission of the k th sub-codeword $k = 1, \dots, K$ and the received signal is given by

$$\mathbf{y}_k = \sqrt{\gamma_k} \mathbf{x}_k + \mathbf{z}_k \quad (3)$$

where \mathbf{z}_k is the vector of zero-mean complex, unitary-variance uncorrelated Gaussian variables (modelling noise). The signal-to-noise ratio (SNR) γ_k defines the channel state information (CSI) which is perfectly known/estimated at the receiver, but unknown to the transmitter. SNR does not change during the transmission of the sub-codeword but varies independently from one sub-codeword to another. This corresponds to a practical scenario where subsequent sub-codewords are not sent in adjacent time instants and, being sufficiently well separated, the realizations of the SNR become—to all practical extent—independent.

¹The receiver can determine if the decoding error occurs using an outer error-check code which causes the transmission overhead which we neglect for simplicity of the analysis.

The channel gains $\sqrt{\gamma}$ are Nakagami- m distributed, so the SNR is characterized by the gamma function (PDF)

$$p(\gamma; m) = \frac{\gamma^{m-1}}{\Gamma(m)} \left(\frac{m}{\bar{\gamma}}\right)^m \exp\left(-\frac{m\gamma}{\bar{\gamma}}\right). \quad (4)$$

where $\bar{\gamma}$ is the average SNR. The cumulative density function of SNR is thus given by

$$F(x; m) = \int_0^x p(\gamma; m) d\gamma = \Gamma(m, mx/\bar{\gamma}) \quad (5)$$

with $\Gamma(m, \gamma) = \frac{1}{\Gamma(m)} \int_0^\gamma x^{m-1} e^{-x} dx$ and $\Gamma(m) = \Gamma(m, \infty)$ are, respectively, the incomplete gamma function and the gamma function.

The coding scheme is revealed to the transmitter, which in the k th transmission implements a maximum likelihood decoding using the observations $\tilde{\mathbf{y}}_k = [\mathbf{y}_1, \dots, \mathbf{y}_k]$.

The system-level implementation of the variable-rate HARQ described above deserves some comments. Namely, we may assume that each transmission contains only one sub-codeword in which case the duration of transmission attempts must vary. This might be a valid approach for a single-user communication where the transmitter and the receiver can negotiate the transmission time for each sub-codeword. On the other hand, it may be a questionable strategy in multi-user communications, where sharing the requirement for a variable-rate transmission with all the users is not practical. It might be possible to assign the resources (time) independently of the varying transmission length but it would lead to the bandwidth loss (sub-codewords shorter than the assigned transmission time slot) or to collisions (sub-codewords longer than the available time).

To avoid such a conceptual difficulty, we assume that the sub-codewords corresponding to different packets are gathered in frames that have the duration of N_F symbols. Such an assumption, also used in [19], [20] allows us to deal with variable-rate codewords to fill up the frame and corresponds to TDMA-type communication, where users are provided with a fixed transmission time (frame). This is shown schematically in Fig. 1. We can easily see that the relative loss due to variable length of the sub-codewords can be made arbitrarily small, provided the number of packets in each frame is sufficiently large.

III. ACHIEVABLE THROUGHPUT

The definition of the throughput we use here follows [2]; according to the *reward-renewal* theorem [21] it is the ratio between the expected number of correctly received bits (after up to

K transmissions) and the expected number of channel uses \overline{N}_s required by the HARQ protocol to deliver the packet (in up to K transmission attempts).

We denote by NACK_k , the event of decoding failure in the k -th transmission and by $f_k = \Pr\{\text{NACK}_1, \dots, \text{NACK}_{k-1}, \text{NACK}_k\}$ – the probability of decoding failure after k transmission attempts. The throughput can be then expressed as [5] [9]

$$\eta_K(\rho_1, \dots, \rho_K) = \frac{1 - f_K}{\rho_1 + \sum_{k=2}^K f_{k-1} \rho_k}. \quad (6)$$

which generalizes the results of [2] to the case of transmission with variable sub-codewords' lengths. Note that f_K has the meaning of “HARQ outage”, that is, the probability of losing the data packet after the HARQ process is terminated.

The formulation (6) is entirely general and depends only on the model of the channel and on the coding/decoding scheme. For example, it was used in [9] for convolutionally coded transmission while [2] used it in independently block-fading channel assuming that capacity-achieving codes are available but under constraint $\rho_k \equiv \rho_1$. Here, we remove this constraint but still follow the approach of [2] that has the virtue of providing limits to any practical coding/decoding scheme. We thus assume that the coding/decoding scheme is “capacity-achieving” in the sense that the transmission is successful if the effective transmission rate is not greater than the accumulated mutual information between the sent and the received signals.² This assumption as well as the way the transmitter/receiver deal with the retransmissions will affect the variables f_k used in (6). Namely, three HARQ schemes are considered:

A. HARQ-I

In HARQ type-I (HARQ-I), after k transmissions, only the most recent received sub-codeword is used for decoding and others are discarded (in [2] this scheme was denoted as ALO). In such a case, the decoding failures are independent of each others and the probability of losing a packet after k transmissions is calculated as [2]

$$f_{I,k} = \prod_{l=1}^k \Pr\{C(\gamma_l) \rho_l < 1\} = \prod_{l=1}^k \nu(\rho_l) \quad (7)$$

²The existence of the codes satisfying this criterion when $N_b \rightarrow \infty$ is discussed, e.g., in [22] [16].

where where $C(\gamma) = \log_2(1 + \gamma)$ is the average mutual information (per channel use) when transmitting with SNR γ and $\nu(\rho) = F(2^{1/\rho} - 1; m)$ is the probability of outage (after a single transmission) when transmitting with redundancy ρ .

The throughput of HARQ-I is then given by

$$\eta_{I,K}(\rho_1, \dots, \rho_K) = \frac{1 - \prod_{k=1}^K \nu(\rho_k)}{\rho_1 + \sum_{k=2}^K \rho_k \prod_{l=1}^{k-1} \nu(\rho_l)} \quad (8)$$

and the optimal throughput is denoted as $\hat{\eta}_{I,K} = \max_{\rho_1, \dots, \rho_K} \eta_{I,K}(\rho_1, \dots, \rho_K)$.

Proposition 1: The maximal throughput of HARQ-I $\hat{\eta}_{I,K}$ is independent of K , i.e., $\hat{\eta}_{I,K} \equiv \hat{\eta}_I = \eta_{I,1}(\hat{\rho}_I)$ where $\hat{\rho}_I = \arg_{\rho} \max \frac{1-\nu(\rho)}{\rho}$, and is yield with fixed-rate HARQ (FR-HARQ-I) $\hat{\rho}_{I,l} = \hat{\rho}_I, l = 1, \dots, K$.

Proof: Since $\hat{\eta}_{I,1} \geq \frac{1-\nu(\rho)}{\rho}$, where the equality hold only for $\rho = \hat{\rho}_I$, we can use $\rho_k \geq \frac{1-\nu(\rho_k)}{\hat{\eta}_{I,1}}$ in (8), which yields the following inequality

$$\begin{aligned} \eta_{I,K}(\rho_1, \dots, \rho_K) &\leq \hat{\eta}_{I,1} \frac{1 - \prod_{k=1}^K \nu(\rho_k)}{1 - \nu(\rho_1) + \sum_{k=2}^K (1 - \nu(\rho_k)) \prod_{l=1}^{k-1} \nu(\rho_l)} \\ &= \hat{\eta}_{I,1} \frac{1 - \prod_{k=1}^K \nu(\rho_k)}{1 - \nu(\rho_1) + \sum_{k=1}^{K-1} \prod_{l=1}^k \nu(\rho_l) - \sum_{k=2}^K \prod_{l=1}^k \nu(\rho_l)} \\ \eta_{I,K}(\rho_1, \dots, \rho_K) &\leq \hat{\eta}_{I,1} = \hat{\eta}_I \end{aligned} \quad (9)$$

thus $\hat{\eta}_I$ is the maximum throughout of VR-HARQ-I, achievable only if $\rho_k = \hat{\rho}_I, k = 1, \dots, K$.

According to Proposition 1, the fixed-rate HARQ-I is optimal so the same sub-codeword may be used for each transmission and the transmitter can apply the RTC transmission scheme defined in Sec. II.

Proposition 1 that is valid for any K may be seen as a generalization of Corrolary 1 in [5] valid for $K \rightarrow \infty$.

B. HARQ-IR

In incremental redundancy HARQ (HARQ-IR) the transmitted sub-codewords are obtained according to IR principle described in Sec. II and the decoding fails in the k -th transmission attempt if the accumulated mutual information is lower than the transmission rate, which yields the following condition [16]

$$f_{IR,k} = \Pr \left\{ \sum_{l=1}^k C(\gamma_l) \rho_l < 1 \right\}. \quad (10)$$

where γ_l is the SNR during l th transmission attempt.

To calculate $f_{\text{IR},k}$ we may proceed as suggested in [2] introducing random variable $v_l = C(\gamma_l) \cdot \rho_l$, $l = 1, \dots, k$ whose PDF can be obtained by definition as $g_l(x) = \ln(2) \cdot p(2^{x/\rho_l} - 1; m) 2^{x/\rho_l} / \rho_l$. This is what will be called the “exact” calculation.

Alternatively, we may approximate v_l by a Gaussian variable, [3], i.e.,

$$g_l(x) \approx \tilde{g}_k(x) = \frac{1}{\sqrt{2\pi}\rho_l\sigma_m} \exp\left(-\frac{(x - \overline{C}_m\rho_l)^2}{2\rho_l^2\sigma_m^2}\right) \quad (11)$$

where

$$\overline{C}_m = \int_0^\infty C(\gamma)p(\gamma;m)d\gamma \quad (12)$$

$$\sigma_m^2 = \int_0^\infty C^2(\gamma)p(\gamma;m)d\gamma - \overline{C}_m^2 \quad (13)$$

are, respectively the mean of $C(\gamma)$ (i.e., the ergodic capacity), and the variance of $C(\gamma)$.

Since $v_l, l = 1, \dots, k$ are independent, $f_{\text{IR},k} = \Pr\left\{\sum_{l=1}^k v_l < 1\right\} = \int_0^1 \overline{g}_k(x)dx$, where $\overline{g}_k(x)$ is a convolution of $g_l(x)$. The latter must be calculated numerically, e.g., via direct/inverse Fourier transform if the exact form of $g_l(x)$ is used, while, applying (11) we obtain a closed-form approximation of (10)

$$f_{\text{IR},k} \approx \tilde{f}_{\text{IR},k} = Q\left(\xi \frac{X_k - 1}{Y_k}\right), \quad (14)$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-t^2/2)dt$, $\xi = \frac{\overline{C}}{\sigma_C}$, $X_k = \sum_{l=1}^k \rho'_l$, $Y_k = \sqrt{\sum_{l=1}^k \rho'^2_l}$, and $\rho' = \rho \cdot \overline{C}$.

Proposition 2: Denoting by $\tilde{\eta}_{\text{IR},K}(\rho_1, \dots, \rho_K)$ the approximation of the throughput obtained using (14) in (6), the following inequality holds

$$\tilde{\eta}_{\text{IR},K}(\rho_1, \dots, \rho_K) \geq \tilde{\eta}_{\text{IR},K}(\rho_1, \dots, \rho_K) = \overline{C} \frac{1 - \tilde{f}_{\text{IR},K}}{\rho'_1 + \sum_{k=2}^K \rho'_k \cdot \tilde{f}_{\text{IR},k-1}} \quad (15)$$

where $\tilde{f}_{\text{IR},k} = Q(\xi(1 - 1/X_k))$.

Proof: The obvious relationship $Y_k^2 \leq X_k^2$ used in (14) yields $\tilde{f}_{\text{IR},k} \leq \check{f}_{\text{IR},k}$. From this inequality and knowing that the throughput decreases monotonically with f_k (if ρ_k are kept constant), we immediately obtain the lower bound (15).

The bound (15) will be useful to optimize the throughput in Sec. IV.

C. HARQ-CHASE

Instead of discarding the packets that were not decoded correctly (as done in HARQ-I), the receiver should take advantage of all received packets and if RTC is employed, the received signals should be weighted by the corresponding SNR and added up. This is known as maximum ratio combining (MRC) or Chase combining [23]. Then, the decoder used the following signal

$$\tilde{\mathbf{y}}_k = \sum_{l=1}^k \sqrt{\gamma_l} \cdot \mathbf{y}'_l \quad (16)$$

where $\mathbf{y}'_k = [\mathbf{y}_k, 0, 0, \dots, 0]$ are zero-padded version of the received signal \mathbf{y}_k . The padding is used to make the notation compact and may be seen as an operation carried out at the receiver thus it does not affects the throughput.

Although, in the case of a fixed-rate HARQ, it was shown to introduce little gain over HARQ-I [2], [13], Chase-combining is the most the receiver can do when RTC is implemented at the transmitter so we deal with this case for completeness of our analysis.

Calculation of the decoding failure probability is slightly more involved in this case.

First, for convenience, we reorder the variables $\gamma_1, \dots, \gamma_k$ so that the corresponding sub-codewords lengths' after reordering are non-decreasing $\rho_{\kappa_1} \leq \rho_{\kappa_2} \leq \dots \leq \rho_{\kappa_k}$, where $\kappa_1, \dots, \kappa_k$ is a permutation of $1, \dots, k$. We emphasize that the reordering is merely a concept simplifying the analysis and not an a priori constraints on the values ρ_k .

As illustrated in Fig. 2, thanks to the reordering we are able to identify k “chunks” of the sub-codewords

$$\begin{aligned} \boldsymbol{\epsilon}_{k,1} &= [x_1, \dots, x_{N_{s,\kappa_1}}], \\ \boldsymbol{\epsilon}_{k,l} &= [x_{N_{s,\kappa_{l-1}}+1}, \dots, x_{N_{s,\kappa_l}}], \quad l = 2, \dots, k, \end{aligned}$$

each with normalized redundancy $\tilde{\rho}_l = \rho_{\kappa_l} - \rho_{\kappa_{l-1}}$ (we set $\rho_{\kappa_0} \equiv 0$), such that all symbols in the chunk $\boldsymbol{\epsilon}_{k,l}$ were transmitted in the transmissions attempts indexed with $\kappa_l, \kappa_{l+1}, \dots, \kappa_k$.

After simple algebra, the combining of the received signals (16) yields

$$\tilde{\mathbf{y}}_k = [\tilde{\mathbf{y}}_{k,1}, \tilde{\mathbf{y}}_{k,2}, \dots, \tilde{\mathbf{y}}_{k,k}] \quad (17)$$

$$\tilde{\mathbf{y}}_{k,l} = \sqrt{\tilde{\gamma}_{k,l}} \left[\sqrt{\tilde{\gamma}_{k,l}} \cdot \boldsymbol{\epsilon}_{k,l} + \boldsymbol{\xi}_{k,l} \right] \quad (18)$$

where

$$\tilde{\gamma}_{k,l} = \sum_{f=l}^k \gamma_{\kappa_f}. \quad (19)$$

is the equivalent SNR for the chunk $\epsilon_{k,l}$ and $\xi_{k,l}$ is a zero-mean, unitary-variance Gaussian vector modelling “equivalent noise” affecting the chunk $\epsilon_{k,l}$.

Since the symbols in the chunks $\epsilon_{k,l}$ are mutually independent, the parts $\tilde{y}_{k,l}$ of the received signal \tilde{y}_k may be seen as the result of transmission of the chunks $\epsilon_{k,l}$ over the channel with SNR $\tilde{\gamma}_{k,l}$. Consequently, Chase combining may be seen as a form of IR transmission with redundancy $\tilde{\rho}_l$ and the probability of the decoding failure is given by

$$f_{\text{CH},k} = \Pr \left\{ \sum_{l=1}^k C(\tilde{\gamma}_{k,l}) \tilde{\rho}_l < 1 \right\} \quad (20)$$

that, in the case of a fixed-rate transmission, boils down to the formula shown in [2]. Namely, since in fixed-rate HARQ-CHASE $\tilde{\rho}_1 = \rho_1$, $\tilde{\rho}_l = 0, l = 2, \dots, k$, and $\tilde{\gamma}_{k,1} = \sum_{l=1}^k \gamma_k$, then the decoding failure is calculated in a closed form

$$f_{\text{CH},k} = \Pr \left\{ C \left(\sum_{l=1}^k \gamma_k \right) \cdot \rho_1 < 1 \right\} = F(2^{1/\rho_1} - 1; m \cdot k). \quad (21)$$

While (20) resembles (10), the equivalent SNRs $\tilde{\gamma}_{k,l}$ appearing in (20) are not independent (unlike in the case of HARQ-IR), so the approach of Sec. III-B, based on the convolution of the individual PDFs cannot be applied and a multidimensional integration over $\gamma_1, \dots, \gamma_k$ is required

$$f_{\text{CH},k} = \int_{\mathcal{D}_k} \prod_{l=1}^k p(\gamma_l; m) d\gamma_1 \dots d\gamma_k \quad (22)$$

where $\mathcal{D}_k = \{\gamma_1, \dots, \gamma_k : \sum_{l=1}^k C(\tilde{\gamma}_{k,l}) \tilde{\rho}_l < 1\}$, so

$$f_{\text{CH},k} = \int_0^{z_1} p(\gamma_1; m) d\gamma_1 \dots \int_0^{z_{k-1}} p(\gamma_{k-1}; m) d\gamma_{k-1} \int_0^{z_k} p(\gamma_k; m) d\gamma_k, \quad (23)$$

where the integration limit for the SNR γ_l depends on the values taken by the SNRs $\gamma_{l+1}, \dots, \gamma_k$

$$z_l \equiv z_l(\gamma_{l+1}, \dots, \gamma_k) = \left[2^{R_l(1 - \sum_{f=l+1}^k \frac{1}{R_l} [C(\tilde{\gamma}_f) - C(\tilde{\gamma}_{f+1})])} - 1 \right] (1 + \tilde{\gamma}_{l+1}). \quad (24)$$

To implement (23) we used the Gauss-Laguerre formulae with 10 (for $m = 1, 2$) or 40 (for $m = \frac{1}{2}$) points in each of k dimensions of \mathcal{D}_k .

The multidimensional calculation was particularly computationally-intensive for $K > 4$ and the results do not seem very relevant beyond this point as virtually all improvement is due to the second transmission attempt.

D. Limiting cases

We know from [2] that for a fixed-rate HARQ-IR

$$\hat{\eta}_{\text{IR},K} \xrightarrow{K \rightarrow \infty} \overline{C}_m \quad (25)$$

where \overline{C}_m is the ergodic capacity, defined in (25).

We also known that, for a given set of ρ_1, \dots, ρ_K , the relationship $f_{\text{I},k} > f_{\text{CH},k} > f_{\text{IR},k}$ holds for all k [2], and since, for the given ρ_1, \dots, ρ_K , the throughput η_K (6) monotonically decreases when f_k increases, we conclude that for $K < \infty$

$$\hat{\eta}_{\text{I},1} = \hat{\eta}_{\text{I},K} < \hat{\eta}_{\text{CH},K} < \hat{\eta}_{\text{IR},K} < \overline{C}. \quad (26)$$

Thus, the throughput of HARQ schemes with fixed-power transmission, operating without knowledge of instantaneous SNR, is lower-bounded by $\hat{\eta}_{\text{I}}$ defined in Sec. III-A and upper-bounded by the ergodic capacity \overline{C} .

The limiting case $K \rightarrow \infty$ is also interesting since, as stated in [5, Lemma 1], when the receiver does not discard packets (as it is the case for HARQ-IR and HARQ-CHASE), the optimal redundancy sequence must be non-increasing, i.e., $\rho_{\text{IR},k} \geq \rho_{\text{IR},k+1}$ and $\rho_{\text{CH},k} \geq \rho_{\text{CH},k+1}$.

IV. OPTIMIZATION

The “design” of the HARQ scheme consists in the maximization of the throughput over the redundancy values ρ_1, \dots, ρ_K . In the case of FR-HARQ the exhaustive search over one-dimensional space is relatively simple. On the other hand, the solutions for VR-HARQ-IR and VR-HARQ-CHASE are more difficult to find as their require a multidimensional optimization.

To maximize (6) we might use a gradient-based method but the initialization of the variables is critical to ensure rapid convergence and to avoid getting trapped far from the global optimum (both - not guaranteed in non-concave functions we deal with, cf. [24, Fig. 1]) , so we used this approach only in VR-HARQ-CHASE where various initializations were tested and the solutions were compared to the random initializations. This was tedious but feasible as it was done only for $K \leq 4$.

In case of VR-HARQ-IR, different approach was adopted: instead of maximizing the throughput $\eta_{\text{IR},K}$ we maximizing the lower bound (15). The problem is greatly simplified since each

term $\check{f}_{\text{IR},k}$ depends uniquely on $X_k = \sum_{l=1}^k \rho'_l$ and the optimization may be written as

$$\max_{\rho_1, \dots, \rho_K} \check{\eta}_{\text{IR}}(\rho_1, \dots, \rho_K) = \max_X \frac{1 - f_K(X)}{V_K(X)} \quad (27)$$

where $f_k(X) = \check{f}_{\text{IR},k} = Q(\xi\sqrt{k}(1 - 1/X))$ and

$$V_k(X) = \min_{\substack{\rho'_1, \dots, \rho'_k: \\ \sum_{l=1}^k \rho'_l = X}} \rho'_1 + \sum_{l=2}^k \rho'_l f_{l-1}(X_{l-1}) \quad (28)$$

$$= \min_{0 \leq \rho'_k \leq X} \min_{\substack{\rho'_1, \dots, \rho'_{k-1}: \\ \sum_{l=1}^{k-1} \rho'_l = X - \rho'_k}} \rho'_1 + \sum_{l=2}^{k-1} \rho'_l f_{l-1}(X_{l-1}) + \rho'_k f_{k-1}(X - \rho_k) \quad (29)$$

$$= \min_{0 \leq \rho \leq X} V_{k-1}(X - \rho) + \rho f_{k-1}(X - \rho). \quad (30)$$

For a given X , the minimization in (30) is done over one variable ($\rho = \rho_k$) provided the results of the minimization $V_{k-1}(X)$ are known for all arguments X . That is, first we solve $V_2(X) = \min_{\rho} \{X - \rho + \rho f_1(X - \rho)\}$, next $V_3(X) = \min_{\rho} \{V_2(X - \rho) + \rho f_1(X - \rho)\}$, etc. This recursive formulation is characteristic of the so-called dynamic programming (DP) [25] whose application for throughput optimization was already suggested in [5]. The direct benefit is that the optimization (27) over K -dimensions is reduced to K , one-dimensional functional optimizations, which greatly simplifies the implementation.

The function $V_k(X)$ is not obtained in the closed-form, so we discretized X using 50-100 points over the domain $X \in (0, k)$, where the bounding of X by k is not restrictive and comes from the heuristic observation that $\rho'_k < 1$, i.e., each rate $R_k = 1/\rho_k$ is greater than the ergodic capacity \bar{C} .

The optimization results are stored as $\rho_k(X) = \arg \min_{\rho} V_{k-1}(X - \rho) + \rho f_{k-1}(X - \rho)$, so once the functions $V_k(X)$ are obtained, we can recover the solution $\hat{\rho}'_k$ that maximizes the bound:

$$\hat{\rho}'_k = \rho_k(\hat{X}_k) \quad (31)$$

where $\hat{X}_K = \arg_X \max \frac{1 - f_K(X)}{V_K(X)}$ and $\hat{X}_{k-1} = \hat{X}_k - \rho_k(\hat{X}_k)$.

We note that while the approximate expressions for $f_{\text{IR},k}$ and $\check{\eta}_{\text{IR},K}$ are used in DP optimization, the throughput values we show in the following are based on the exact calculation of $f_{\text{IR},k}$. We also verified that using the DP-based results as the initialization to the gradient-based optimization yields practically the same values of the throughput as those we show.

V. NUMERICAL RESULTS

The optimized throughput of fixed- and variable-rate HARQ-IR is shown in Fig. 3 for $K = 2, 4, 8$ for Rayleigh fading channel (i.e. with $m = 1$), where the gain due to variable-rate transmission is particularly notable for HARQ-IR while it is very slight when considering HARQ-CHASE, which at best (with $K = 4$) equals the performance of FR-HARQ-IR with $K = 2$.

The gain in terms of throughput offered by VR-HARQ-IR is particularly clear for $K = 2$ and to complement the results of Fig. 3, we evaluate and show in Fig. 4 the “residual throughput”

$$\chi = 1 - \frac{\eta_{\text{IR}}}{\overline{C}} \quad (32)$$

i.e., the relative gap between the throughput attained with up to K transmissions and the maximum achievable throughput (ergodic capacity). The relative gain of the VR-HARQ with respect to FR-HARQ remain roughly constant for all K but of course the absolute difference diminishes – as expected – with K since, asymptotically both schemes are equivalent. These gains are also more notable when increasing m . The “saturation” of the throughput of HARQ-CHASE scheme is also clearly shown.

In Fig. 5 we show the normalized redundancy $\rho'_k = \rho_k \cdot \overline{C}$ directly proportional to the subcodewords' lengths $N_{s,k}$ (inversely proportional to the transmission rates R_k). We observe that the first transmission attempt of VR-HARQ-IR is carried out with the rate $R_1 = 1/\rho_1$ close to \overline{C} , while the rates of subsequent transmissions increase (i.e., the subcodewords are shorter) and decrease again for k approaching K . This relationship holds for all $\overline{\gamma}$ and m and may be observed in the IR and CHASE schemes. In Fig. 6, we reproduce similar results for VR-HARQ-IR and FR-HARQ-IR but for different values of K . The same “profile” of the redundancy is obtained for all K and we may also appreciate that the values of ρ'_k are decreasing with K , which is consistent with the optimal behaviour for $K \rightarrow \infty$, when the optimal sequence of ρ_k should be non-increasing [5, Lemma 1]. For the FR-HARQ-IR, we observe that ρ'_1 decreases with K . Recall that, according to the proof in [2, Appendix C], when $K \rightarrow \infty$ the throughput-maximizing redundancy $\rho'_k = \rho_1 \overline{C}$ should tend to $\frac{1}{K}$.

The decreasing-increasing behaviour of the values ρ'_k can be interpreted from (6) combining the results of Fig. 6 with those in Fig. 7 showing the values of the decoding failure $f_k, k = 1, \dots, K$. Namely, as we strive to make η_K approach closely \overline{C} , from (6) we conclude that redundancy/rate should be allocated so that $\overline{C} \cdot (\rho_1 + \sum_{k=2}^K \rho_k \cdot f_{k-1}) = \rho'_1 + \sum_{k=2}^K \rho'_k \cdot f_{k-1}$ grown to be as close

as possible to $1 - f_K \approx 1$. Immediately we conclude that we have to use $\rho'_1 < 1$ (transmission rate $R_1 > \bar{C}$) but the behaviour of optimal values $\rho'_k, k > 1$ depends on how the values f_k evolve with ρ'_k .

In the particular case of $K \rightarrow \infty$, as long as the receiver “accumulates” the redundancy, the optimal values ρ'_k should be decreasing with k [5]³. Thus, the fact that ρ'_k increases with k (here: for $k > 2$) is due to the truncation (finite K) and reflects the fact that not only the denominator of (6) should be minimized but also we have to guarantee that the value of f_K remains small.

Also, since f_k decreases much faster in HARQ-IR than it does in HARQ-CHASE (due to lack of additional information coded symbols conveyed in the subsequent transmission attempts), the optimal values of $\rho'_k, k > 1$ can be smaller for VR-HARQ-IR than they are for VR-HARQ-CHASE so, as shown in Fig. 5 the variation of the redundancy is less pronounced.

In Fig. 7 we can also observe that for sufficiently large K ($K \geq 4$), the probability of outage $f_{\text{IR},K}$ in VR-HARQ-IR is smaller than in the case of FR-HARQ-IR. For other values of m and γ the same property was consistently observed which is another clear advantage of VR-HARQ-IR over FR-HARQ-IR.

Another consequence of using short sub-codewords for all transmission attempts in FR-HARQ-IR is that the mutual information accumulates “slowly” with the retransmissions. Consequently, the failures in the initial transmissions occur more likely than in the VR-HARQ-IR, where the first transmission is done with the rate R_1 close to \bar{C} . This impacts the average number of transmissions which we calculate as

$$K_{\text{avg}} = 1 + \sum_{k=1}^{K-1} f_k \quad (33)$$

and show in Fig. 8.

We can appreciate that when the number of transmission K grows, the average number of transmissions K_{avg} increases as well but is significantly greater for fixed-rate HARQ-IR: it practically doubles for $K = 8$ and $\gamma = 30\text{dB}$. Since the average number of transmissions is related to the packet delivery delay (as retransmission can be done only in separate frames), VR-HARQ-IR –besides the increased throughput– offers an additional advantage over FR-HARQ-IR.

³Remember that for HARQ-I, i.e., when the receiver discards the redundancy of past transmission attempts, the optimal solution is $\rho_k \equiv \rho_1$

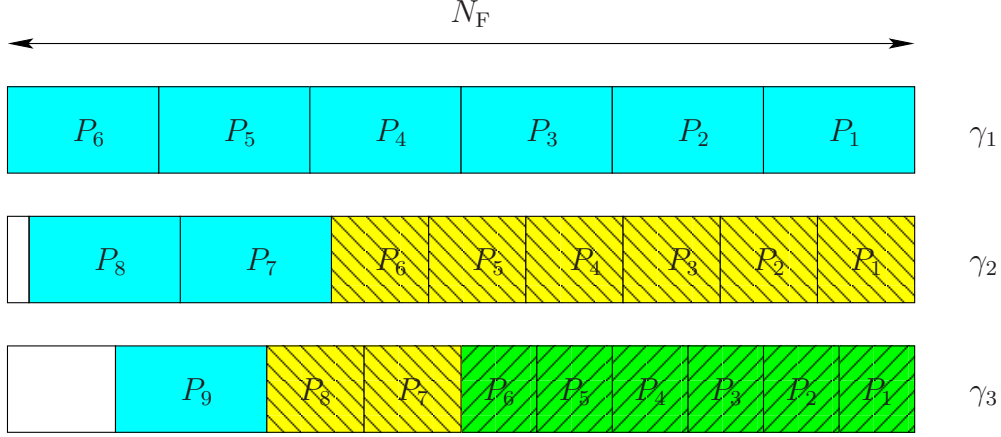


Figure 1. Example of the structure of three frames sent over channels with corresponding SNRs γ_1 , γ_2 , and γ_3 when delivering data packets denoted by $P_l, l = 1, \dots, 9$. The subcodewords having different lengths are identified with different colors and patterns. The first frame is filled up with subcodewords of length $N_{s,1}$ (thus, in our example, $N_F = 6N_{s,1}$) corresponding to the packets $P_1 - P_6$. When transmitting this frame with SNR γ_1 , we assume $C(\gamma_1)\rho_1 < 1$, consequently, the decoder fails to decode the message in the packets $P_1 - P_6$ and a NACK messages are sent to the transmitter. The next frame contains thus six subcodewords of length $N_{s,2}$ each carrying the redundancy for the undelivered packets and since, here, $N_{s,1} > N_{s,2}$, the “empty” space is filled with two subcodewords of the length $N_{s,1}$ corresponding of the packets P_7 and P_8 that are ready for transmission. None of the packets is decoded after the transmission of the second frame so, again, six sub-codewords of length $N_{s,3}$, corresponding to the packets $P_1 - P_6$ are sent as well as the sub-codewords of length $N_{s,2}$ corresponding to the packets P_7 and P_8 . The residual time is filled with the sub-codeword corresponding to the packet P_9 . Note, that the relative loss due to unshaded/unfilled space can be made arbitrarily small loading the frame with many sub-codewords.

VI. CONCLUSIONS

In this paper we have analyzed HARQ with incremental redundancy (HARQ-IR) for transmissions over block-fading channels. We have proposed an efficient method to allocate the optimal rates and have demonstrated that the variable-rate HARQ-IR provides gains over the fixed-rate HARQ-IR in terms of increased throughput, lower outage, and decreased average number of transmissions.

ACKNOWLEDGMENT

The authors thank Dr. M. Benjillali (INPT, Rabat, Morocco) for his critical reading and Prof. J. Benesty (INRS-EMT, Montreal) for the suggestions leading to the simplification of the outage calculation in Sec. III-B.

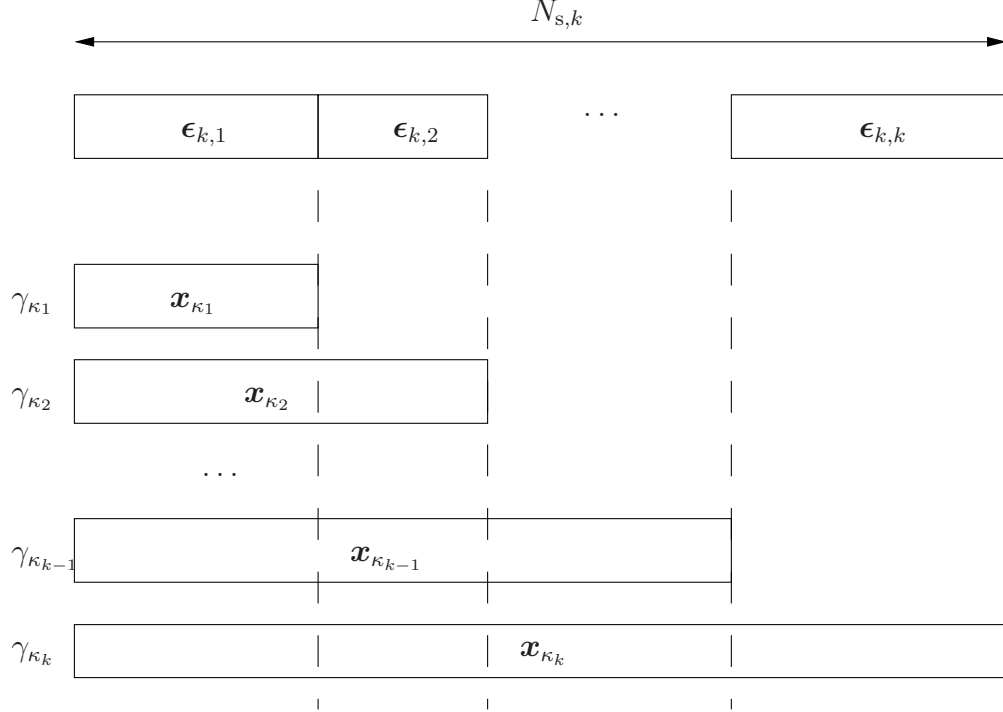


Figure 2. In RTC, the parts of the transmitted sub-codewords that contain the same symbols are identified as “chunks” $\epsilon_{k,l}$; symbols in each chunk experience the same equivalent SNR $\tilde{\gamma}_{k,l} = \sum_{f=l}^k \gamma_{\kappa_f}$.

REFERENCES

- [1] K. Brayer, “Error control techniques using binary symbol burst codes,” *IEEE Trans. Commun.*, vol. 16, no. 2, pp. 199–214, Apr. 1968.
- [2] G. Caire and D. Tuninetti, “The throughput of hybrid-ARQ protocols for the Gaussian collision channel,” *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.
- [3] P. Wu and N. Jindal, “Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis,” *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, Apr. 2010.
- [4] D. Tuninetti, “On the benefits of partial channel state information for repetition protocols in block fading channels,” *CoRR*, vol. abs/1102.4085, 2011.
- [5] E. Visotsky, V. Tripathi, and M. Honig, “Optimum ARQ design: a dynamic programming approach,” in *Proc. IEEE International Symposium on Information Theory*, Jun. 2003, p. 451.
- [6] E. Uhlemann, L. K. Rasmussen, A. Grant, and P.-A. Wiberg, “Optimal incremental-redundancy strategy for type-II hybrid ARQ,” in *Proc. IEEE International Symposium on Information Theory*, 2003, p. 448.
- [7] J.-F. Cheng, Y.-P. Wang, and S. Parkvall, “Adaptive incremental redundancy,” in *IEEE Veh. Tech. Conf.*, Orlando, Florida, USA, Oct. 2003, pp. 737–741.
- [8] R. Liu, P. Spasojevic, and E. Soljanin, “On the role of puncturing in hybrid ARQ schemes,” in *Proc. IEEE International Symposium on Information Theory*, Jun. 2003, p. 449.

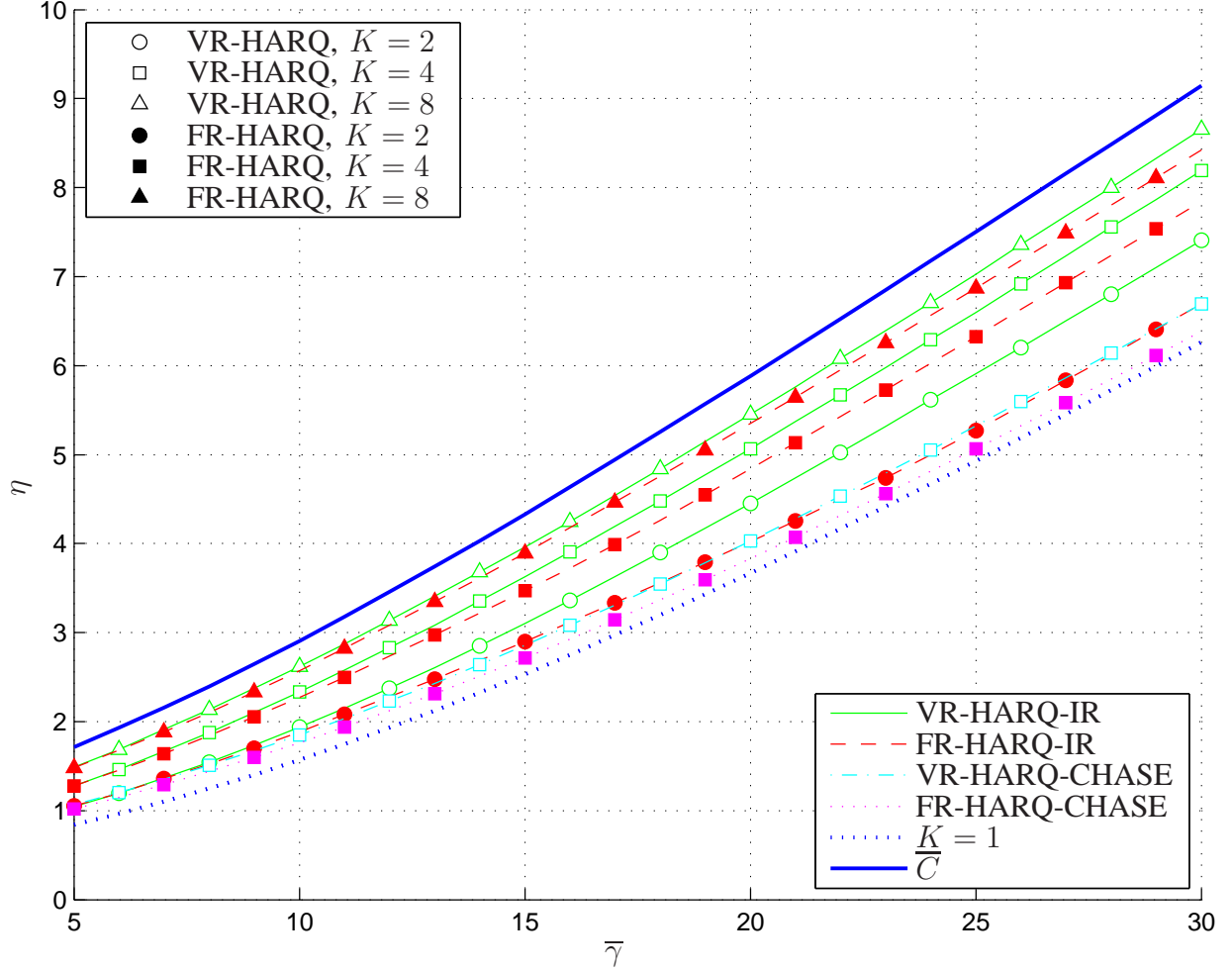


Figure 3. Throughput in a block-fading Rayleigh channel for VR-HARQ-IR (solid, green line), FR-HARQ-IR (dashed, red line) and $K = 2, 4, 8$ as well as VR-HARQ-CHASE (dot-dashed, cyan line) and FR-HARQ-CHASE (dotted, magenta line) shown only for $K = 4$. The upper bound (thick, solid, blue line) corresponds to the ergodic capacity \bar{C} while the lower one (dotted, blue line) corresponds to transmission without HARQ ($K = 1$). That results of VR-HARQ-CHASE ($K = 4$) and VR-HARQ-IR ($K = 2$) are practically superimposed, as well as are those of FR-HARQ-CHASE ($K = 4$) and of transmission without HARQ.

- [9] E. Visotsky, Y. Sun, V. Tripathi, M. Honig, and R. Peterson, "Reliability-based incremental redundancy with convolutional codes," *IEEE Trans. Commun.*, vol. 53, no. 6, pp. 987 – 997, Jun. 2005.
- [10] N. Gopalakrishnan and S. Gelfand, "Rate selection algorithms for IR hybrid ARQ," in *2008 IEEE Sarnoff Symposium*, Princeton, NJ, USA, Apr. 2008, pp. 1–6.
- [11] S. M. Kim, W. Choi, T. W. Ban, and D. K. Sung, "Optimal rate adaptation for hybrid ARQ in time-correlated Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, pp. 968 –979, Mar. 2011.

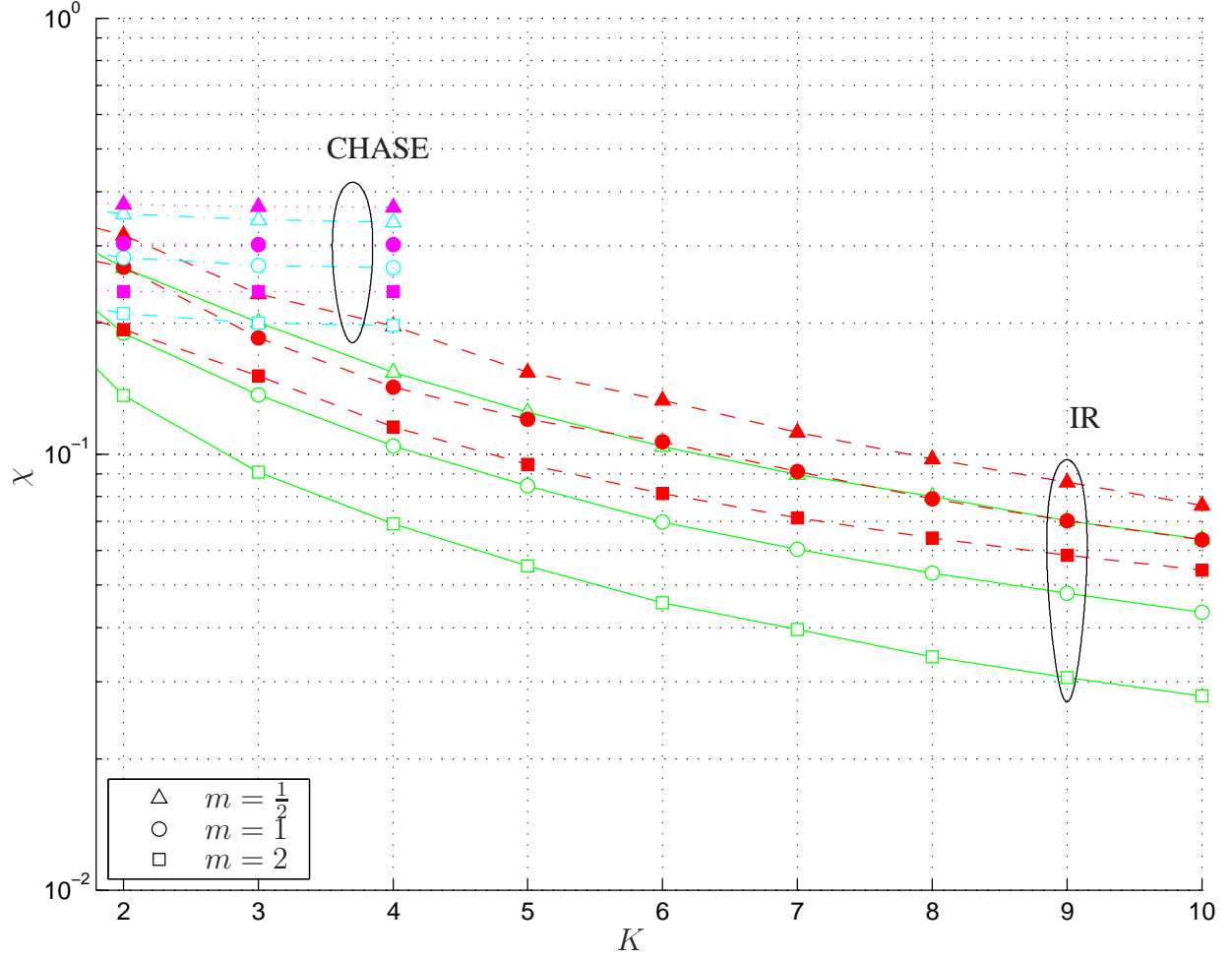


Figure 4. Residual throughput $\xi = 1 - \eta/\overline{C}$ for VR-HARQ-IR (solid, green line), FR-HARQ-IR (dashed, red line), FR-HARQ-CHASE (dotted, blue line), and VR-HARQ-CHASE (dashed-dotted, magenta line) is shown for varying K and Nakagami- m fading with $m = \frac{1}{2}, 1, 2$ and $\overline{\gamma} = 30\text{dB}$. For FR/VR-HARQ-CHASE the results up to $K = 4$ are shown due to high computation load of the throughput calculation.

- [12] C. Shen, T. Liu, and M. Fitz, "On the average rate performance of hybrid-ARQ in quasi-static fading channels," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3339–3352, Nov. 2009.
- [13] D. Tuninetti, "Transmitter channel state information and repetition protocols in block fading channels," in *IEEE Information Theory Workshop, ITW '07*, California, USA, Sep. 2007, pp. 505–510.
- [14] H. Gamal, G. Caire, and M. Damen, "The MIMO ARQ channel: Diversity–multiplexing–delay tradeoff," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3601–3621, Aug. 2006.
- [15] K. D. Nguyen, L. K. Rasmussen, A. G. i Fabregas, and N. Letzepis, "MIMO ARQ with multi-bit feedback: Outage analysis," *CoRR*, vol. abs/1006.1162v2, 2010.
- [16] N. Gopalakrishnan and S. Gelfand, "Achievable rates for adaptive IR hybrid ARQ," in *2008 IEEE Sarnoff Symposium*,

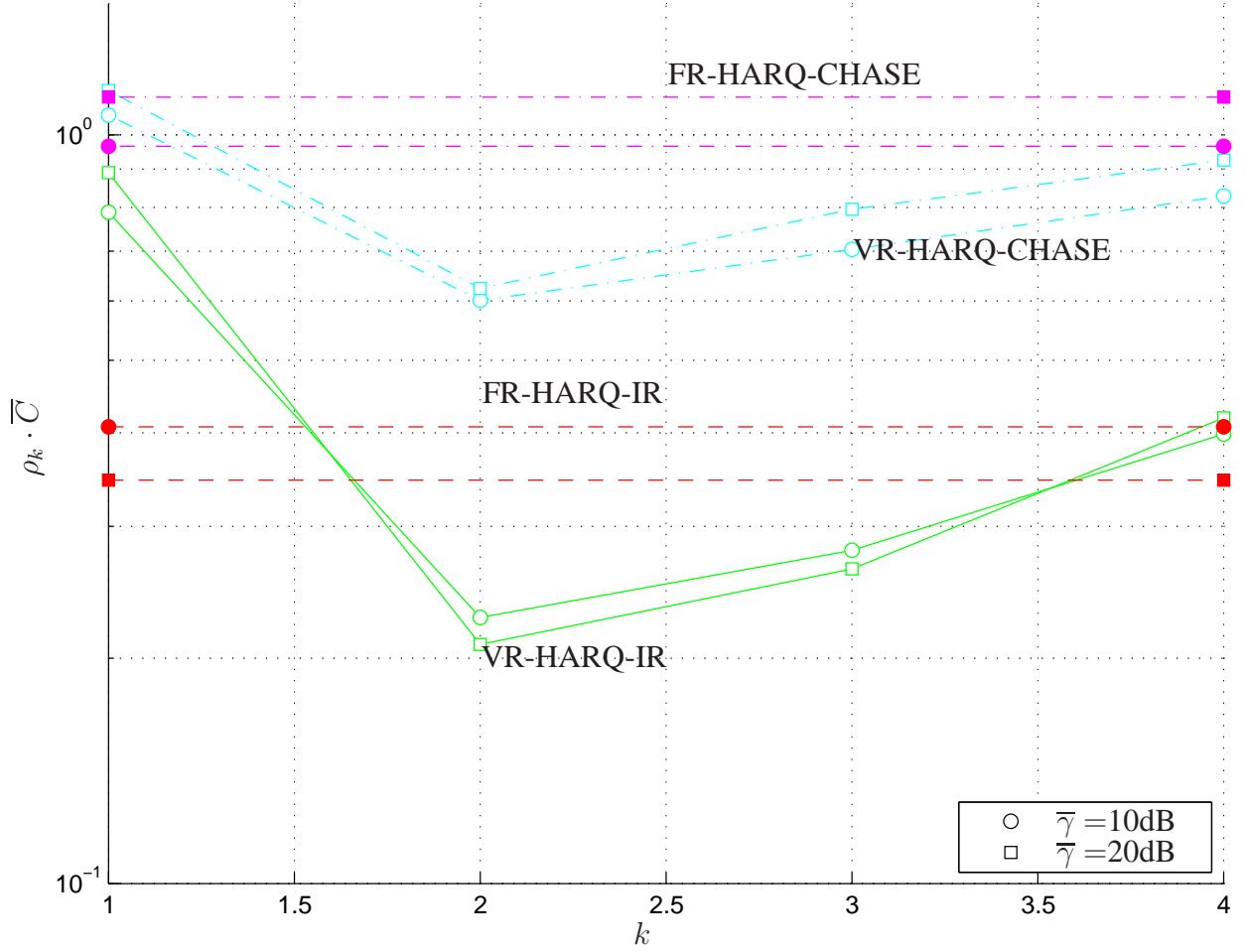


Figure 5. Throughput-maximizing normalized redundancy $\rho'_k = \rho \cdot \overline{C}$, $k = 1, \dots, K$ ($K = 4$) for VR-HARQ-IR (solid, green line), FR-HARQ-IR (dashed, red line), VR-HARQ-CHASE (dashed-dotted, cyan line), and FR-HARQ-CHASE (dashed-dotted, magenta line); for $\overline{\gamma} = 10\text{dB}$ and $\overline{\gamma} = 20\text{dB}$.

Apr. 2008, pp. 1–6.

- [17] T. Kim and M. Skoglund, “On the expected rate of slowly fading channels with quantized side information,” *IEEE Trans. Commun.*, vol. 55, no. 4, pp. 820–829, Apr. 2007.
- [18] A. J. Goldsmith and P. Varaiya, “Capacity of fading channels with channel side information,” *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986–1992, 1997.
- [19] Q. Liu, S. Zhou, and G. B. Giannakis, “Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sep. 2004.
- [20] X. Wang, Q. Liu, and G. Giannakis, “Analyzing and optimizing adaptive modulation coding jointly with ARQ for QoS-guaranteed traffic,” *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 710–720, Mar. 2007.
- [21] M. Zorzi and R. Rao, “On the use of renewal theory in the analysis of ARQ protocols,” *IEEE Trans. Commun.*, vol. 44,

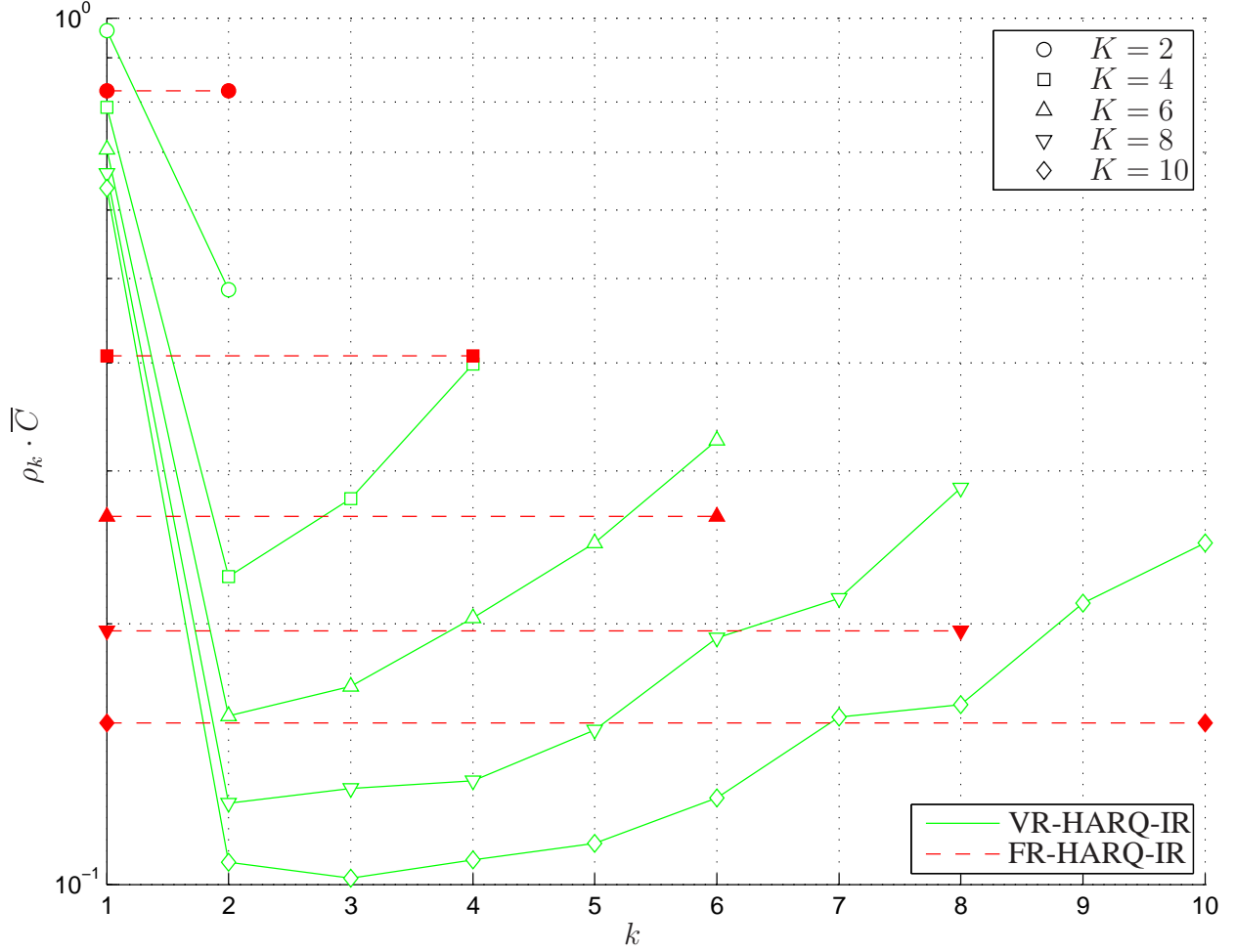


Figure 6. Optimal redundancy $\rho'_k = \rho \cdot \bar{C}, k = 1, \dots, K$ for VR-HARQ-IR (solid, green line) and FR-HARQ-IR (dashed, red line); $m = 1, \bar{\gamma} = 10\text{dB}$.

no. 9, pp. 1077–1081, Sep 1996.

[22] E. Malkamaki and H. Leib, “Coded diversity on block-fading channels,” *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 771–781, Feb. 1999.

[23] J. Cheng, “Coding performance of hybrid ARQ schemes,” *IEEE Trans. Commun.*, vol. 54, no. 6, pp. 1017–1029, Jun. 2006.

[24] L. Szczecinski, C. Correa, and L. Ahumada, “Variable-rate transmission for incremental redundancy hybrid ARQ,” in *IEEE Global Telecommunications Conference, GLOBECOM 2010*, Dec. 2010.

[25] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2005, vol. 1.

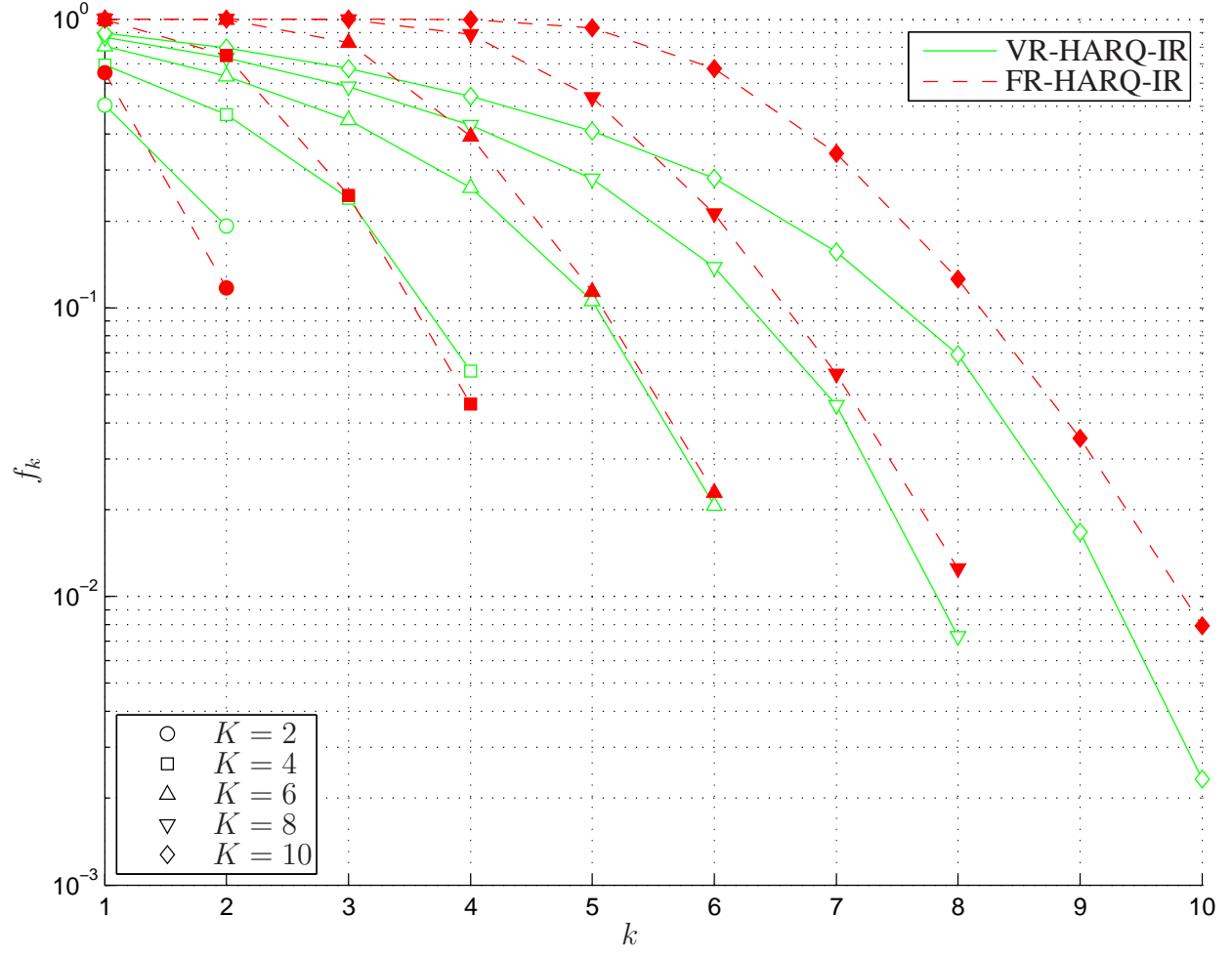


Figure 7. Outage values $f_k, k = 1, \dots, K$ for VR-HARQ-IR (solid, green line) and FR-HARQ-IR (dashed, red line); $m = 1$, $\bar{\gamma} = 10\text{dB}$.

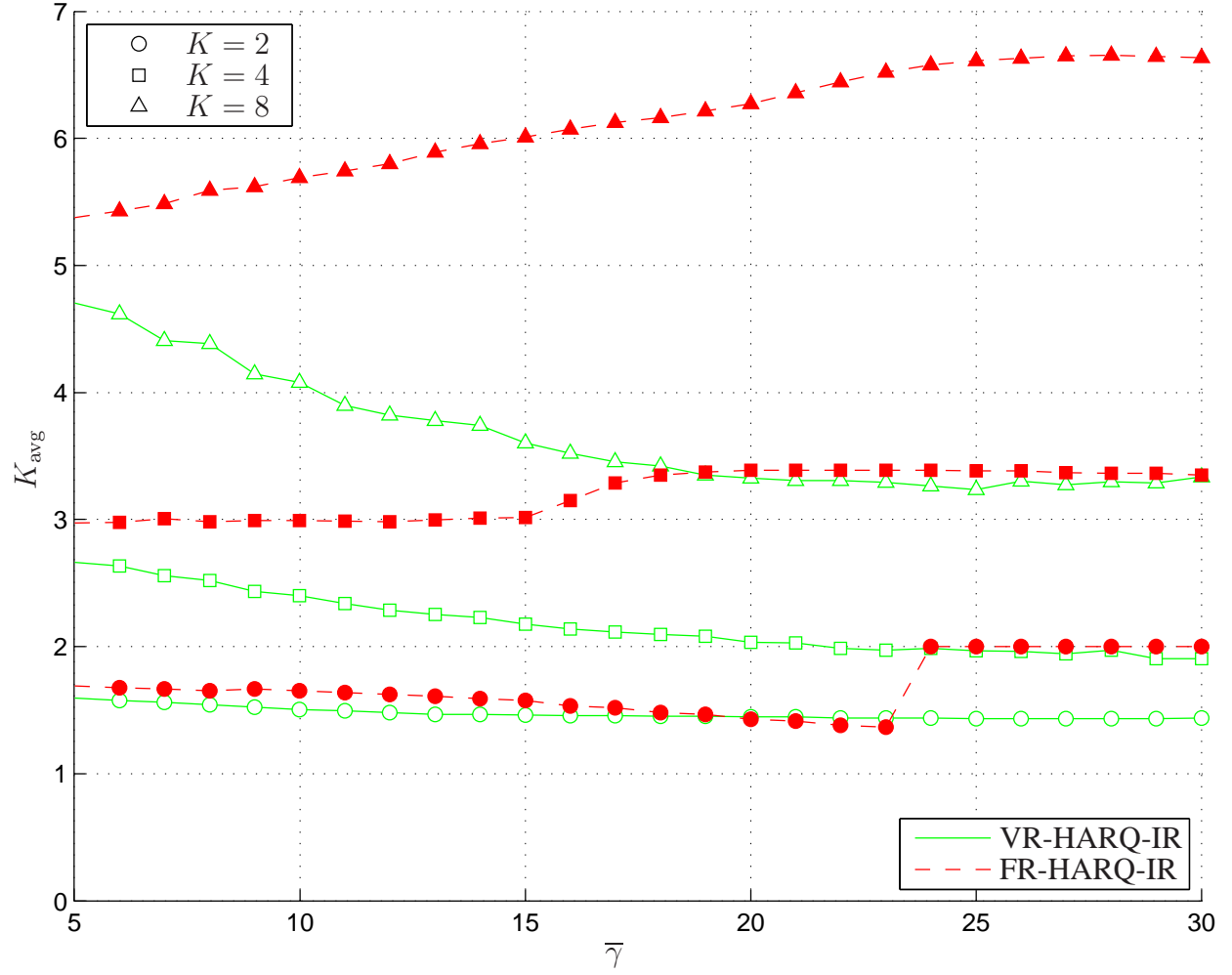


Figure 8. Average number of transmissions K_{avg} for VR-HARQ-IR (solid, green line) and FR-HARQ-IR (dashed, red line); $m = 1$.